SUPPLEMENTARY NOTES

Supplementary Note 1: Scripture: a statistical method for *ab initio* reconstruction of a mammalian transcriptome

We developed Scripture, a genome-guided method to reconstruct the transcriptome using only an RNA-Seq dataset and an (unannotated) reference genome sequence. Scripture consists of six steps (Fig. 1): (1) We use reads uniquely aligned to the genome, including those with gapped alignments spanning exon-exon junctions ('aligned spliced reads')[13] (Fig. 1c); (2) From the aligned spliced reads, we construct a *connectivity graph* representing spliced connections between base pairs in the genome (Fig. 1d); (3) Using all spliced and non-spliced (contiguous) read data, we use a statistical segmentation approach[4] to traverse the connectivity graph and identify significant paths (Fig. 1e); (4) From the paths, we construct a *transcript graph* connecting each exon in the transcript (Fig. 1f); (5) We augment the transcript graph with connections based on paired-end reads and their distance constraints, allowing us to join transcripts or remove unlikely isoforms (Fig. 1g); and (6) We generate a catalogue of transcripts defined by the transcript graph. We discuss each of these steps in detail below.

We first map our reads to the genome, using a gapped aligner[13] that efficiently handles reads that span splicing junctions (**Fig. 1a**). This step is critical since ~30% of 76 base reads are expected on average to span an exon-exon junction (**Methods**). Furthermore, 'spliced' reads provide direct information on the location of splice junctions within the transcript.

We next use only the spliced reads to infer a *connectivity graph* across the genome, where each base in the genome is connected to those bases in the genome that are its immediate neighbors either in the genomic sequence itself or within a spliced read (**Fig. 1d**). Furthermore,

Nature Biotechnology: doi: 10.1038/nbt.1633

we use agreement with splicing motifs at each putative junction in the graph to orient the connection (edge) in the connectivity graph[9, 13] (**Fig. 1d, Methods**).

To infer transcripts, we apply a statistical segmentation approach that identifies significantly enriched paths in the connectivity graph using both spliced and non-spliced reads (**Fig 1e**). Briefly, our segmentation approach identifies regions of mapped read enrichment compared to the genomic background. This is done by scoring a sliding window using a test statistic for each region, computing a threshold for genome-wide significance, and using the significant windows to define intervals (**Methods**). To define intervals, we scan short windows to identify consecutive coverage blocks that have a read coverage scoring above the genome-wide significance threshold we computed. This approach is based on our successful method for identification of chromatin modified regions in genomes[4], but is applied here to the *connectivity graph* rather than to the linear genome.

The result is a set of statistically significant directed *transcript graphs* (**Fig 1f**), each representing one or more splice isoforms of a transcript. Each node in a transcript graph is an exon and each edge is a splice junction. A path through the graph from an exon with no incoming edges (first exon) to an exon with no outgoing edges (last exon) represents one isoform of the gene. Since each graph is directed, all multi-exonic paths are oriented (*i.e.* strand-specific, **Fig. 1e**). Alternative spliced isoforms are identified by considering all possible paths in the transcript graph; since this number may be large and represent spurious paths, we refine it in the next step.

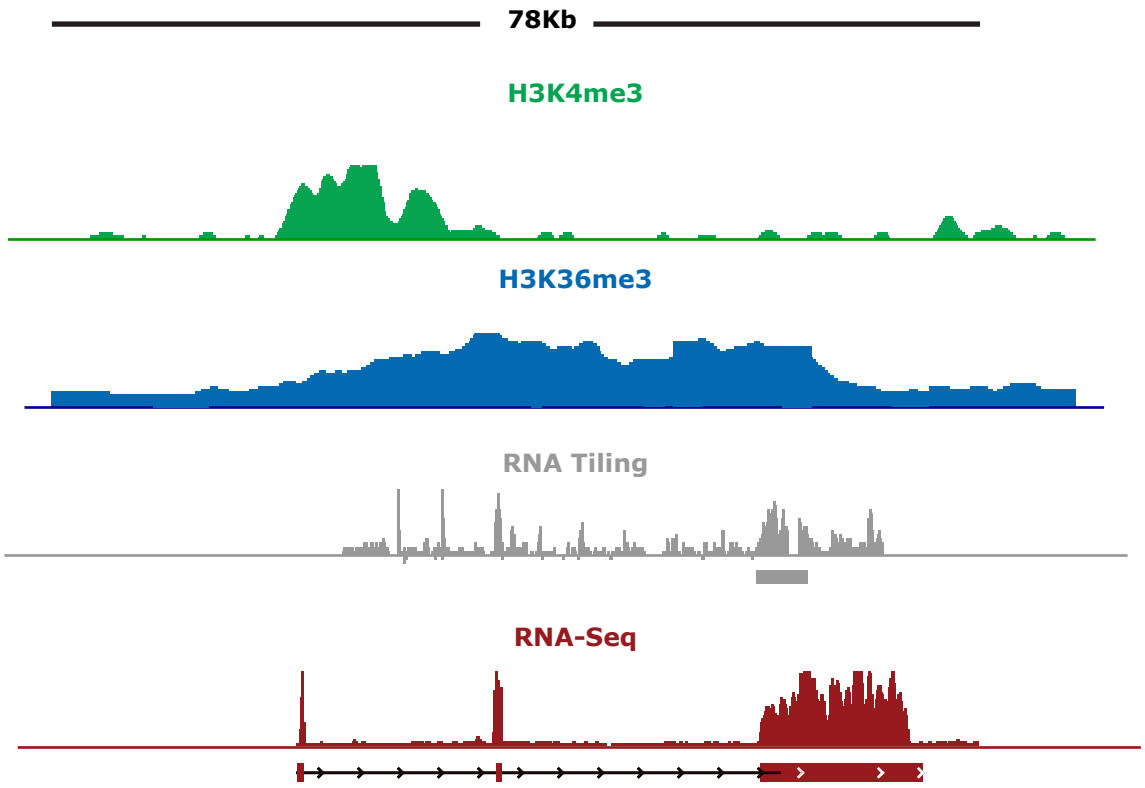
Nature Biotechnology: doi: 10.1038/nbt.1633

Supplementary Note 2: discrepancies between lincRNAs reconstructed in RNA-seq and chromatin-defined lincRNA loci

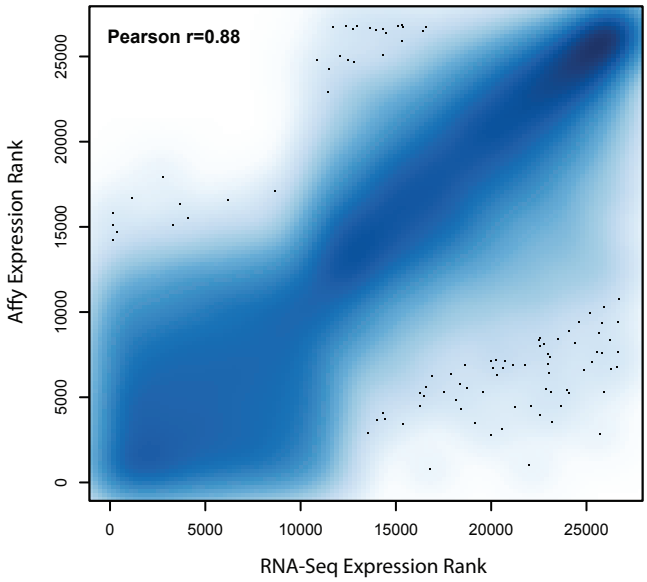
There were few discrepancies between the lincRNAs defined in our previous study in ESC and those reconstructed by Scripture from ESC RNA-Seq. In 11 cases we identified a single reconstructed lincRNA structure that spans across multiple K4-K36 regions, and in 55 cases we it identified two distinct lincRNAs in opposite orientations at a single K4-K36 locus that corresponded to two distinct lincRNA gene structures in opposite orientations. These discrepancies are likely due to the lower resolution of our chromatin maps compared with the base-pair resolution of our transcript maps.

We failed to reconstruct transcripts for a remaining 67 lincRNAs that had been previously identified based on K4-K36 domains in mESC. These lincRNA genes may not have been reconstructed because they are either expressed at lower levels, are single exons, are false positives of our chromatin signature, or are false negatives of the reconstruction approach. For example, 30 of those our previously identified K4-K36 domains are now reconstructed as likely connected to a new isoform of a neighbouring protein coding transcript and thus are no longer counted as lincRNAs in our refined catalogue. The principal reason we miss the remaining 37 cases K4-K36 domains is low expression levels. Nonetheless, 25 of those (67%) of these remaining lincRNA loci (25 lincRNAs) are significantly enriched for reads and likely transcribed (average of 0.76 reads/bp compared to expected 0.01 reads/bp, nominal $p<0.001$, random permutation of reads against size matched random regions). This is consistent with these loci being transcribed. With higher coverage, it should be possible to reconstruct them.

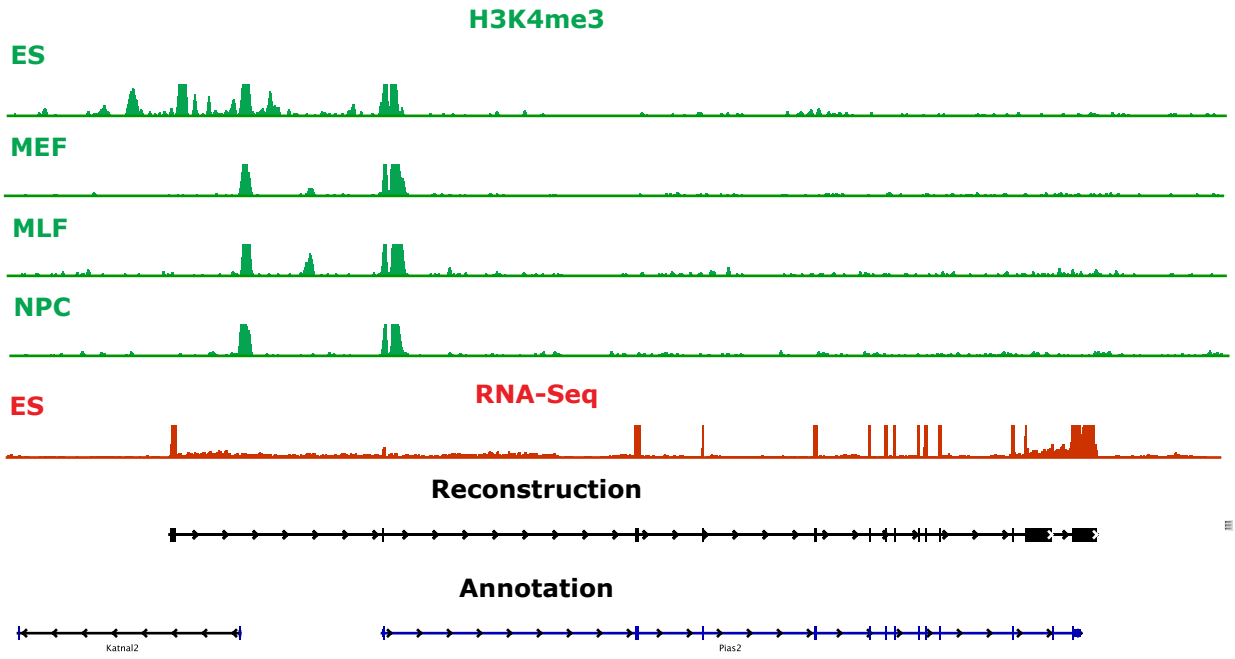
SUPPLEMENTARY FIGURES



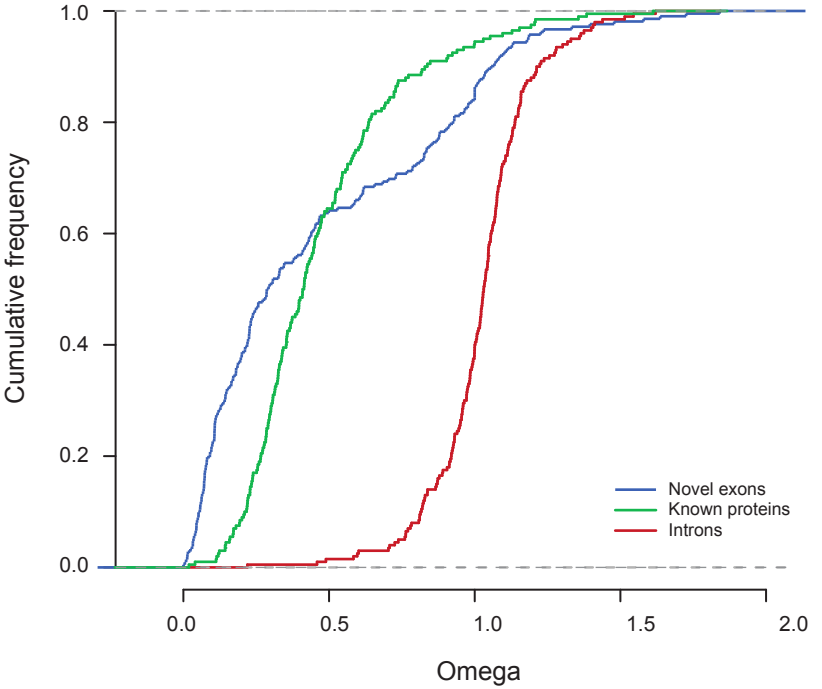
Supplementary Figure 1. Comparison of RNAseq and tiling array gene structures. Shown is a genomic locus for an example lincRNA. Shown are K4me3 (green) and K36me3 (blue) chromatin marks, hybridization signal from a tiling array (grey), RNA-Seq read coverage (red) and Scripture reconstructions. The reconstruction is at substantially higher resolution than the tiling array information, and it identified exons and connectivity missed by the tiling arrays.



Supplementary Figure 2. Correlation in expression levels between our RNA-Seq sample and Affymetrix arrays. Shown is a density scatter plot, showing the estimated expression ranks for annotated protein coding genes from RNA-seq data (by RPKM) and Affymetrix arrays (by MAS 5.0 estimation). Darker shades indicate more genes and lighter shades indicated fewer genes.

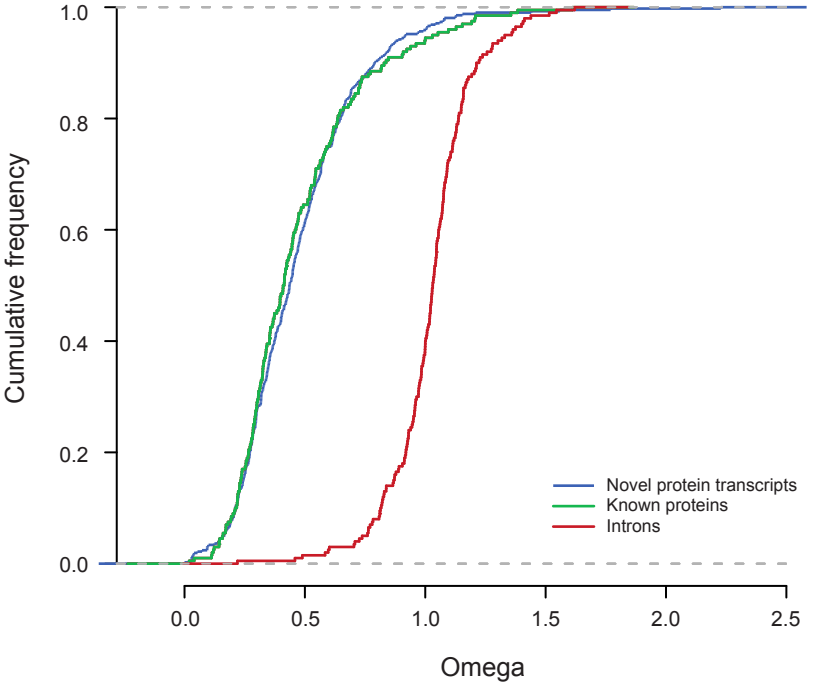


Supplementary Figure 3. H3K4me3 support for an alternative 5' end. Shown is the genomic locus at chr18:77,304,429-77,392,539, along with known annotated transcripts (blue, bottom), the reconstructed transcript from ESC (black) with an alternative external 5' end, and K4me3 data (green tracks, top) for four cell types. The alternative 5' end in ESC is associated with an ESC-specific K4me3 mark.



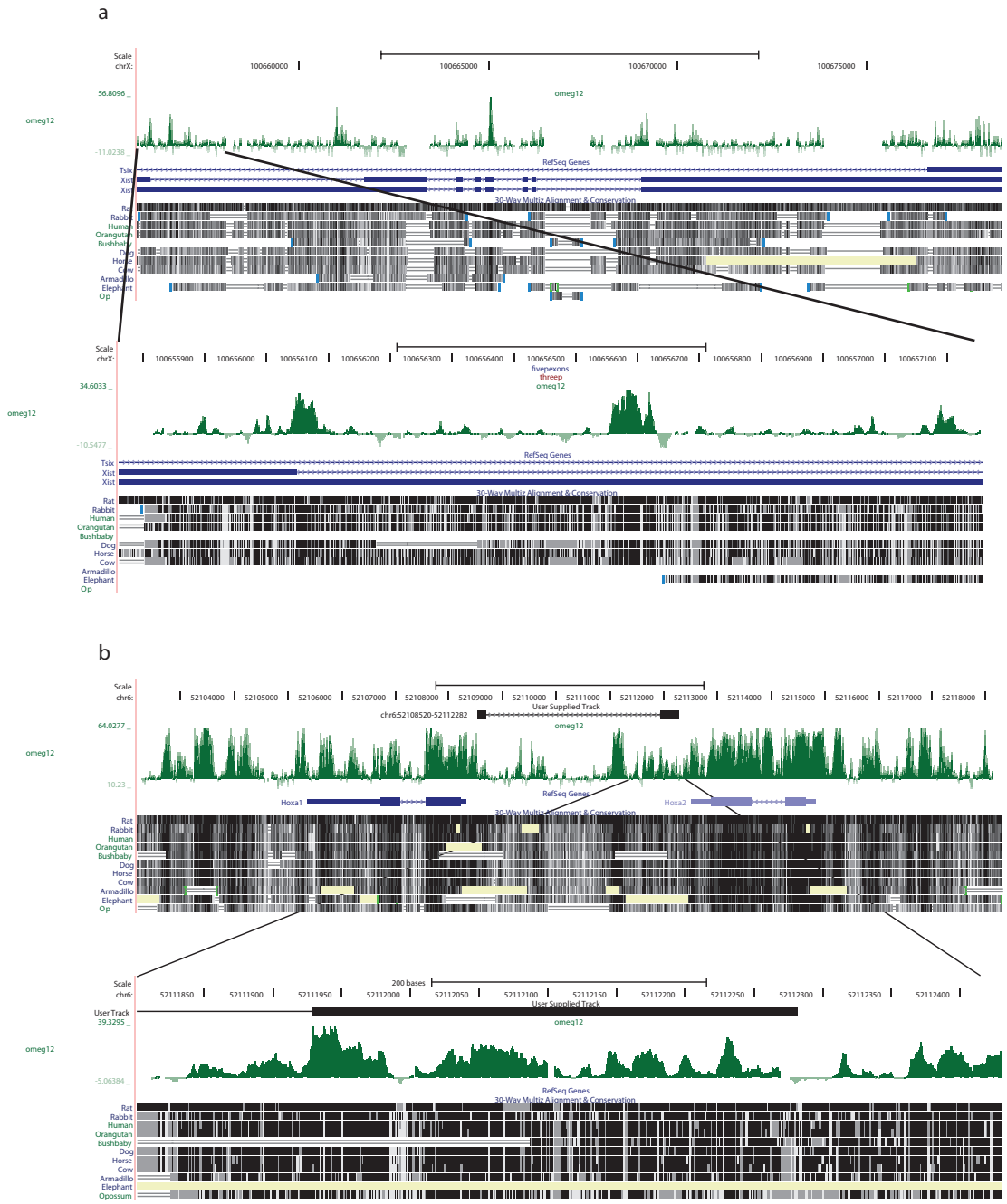
Supplementary Figure 4. Conservation of novel protein coding exons. Shown is the cumulative distribution of sequence conservation across 29 mammals (omega) for novel protein coding exons (blue), annotated protein coding exons (green), and introns (red). The novel exons are conserved at comparable levels to annotated ones.

Nature Biotechnology: doi: 10.1038/nbt.1633



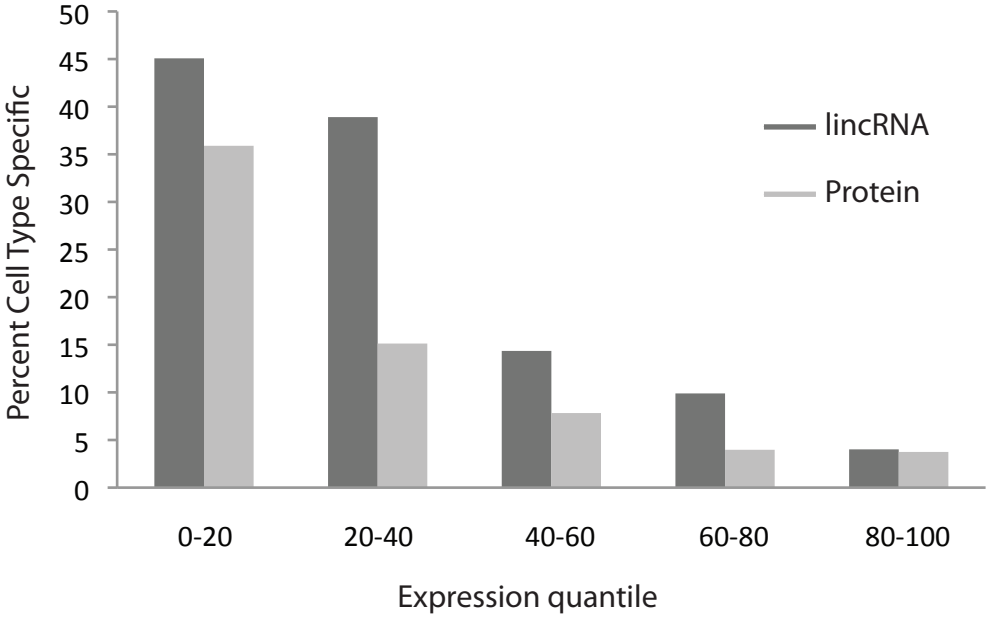
Supplementary Figure 5. Conservation of novel protein coding genes. Shown is the cumulative distribution of sequence conservation across mammals (omega) for exons from novel protein coding genes (blue), annotated protein coding genes (green), and introns (red). The novel genes are conserved at comparable levels to annotated ones.

Nature Biotechnology: doi: 10.1038/nbt.1633



Supplementary Figure 6. Patches of conservation in lincRNAs. Shown are two examples of lincRNAs and the conservation patterns across their exons. In each panel; shown is the relevant genomic locus (top) with a zoom in (bottom), listing omega (top, green), RefSeq gene anontations (middle, blue tracks), and 30-way multiz alignment tracks (bottom, black). (a) Known functional lincRNA, XIST. Zoom in on its 5' exon illustrates the conservation patterns within the exon. (b) A novel lincRNA. Zoom in on one of its exons illustrates the conservation patterns within the exons, comparable to those in (a).

Nature Biotechnology: doi: 10.1038/nbt.1633



Supplementary Figure 7. LincRNAs are more cell type specific across different expression levels. Shown are the fractions of ES specific transcripts for lincRNAs (dark bars) and protein coding genes (light bars) at different ESC expression quantiles. In each quantile, the fraction of lincRNAs with ES specific expression is higher than the fraction of protein coding transcripts with ES specific expression.